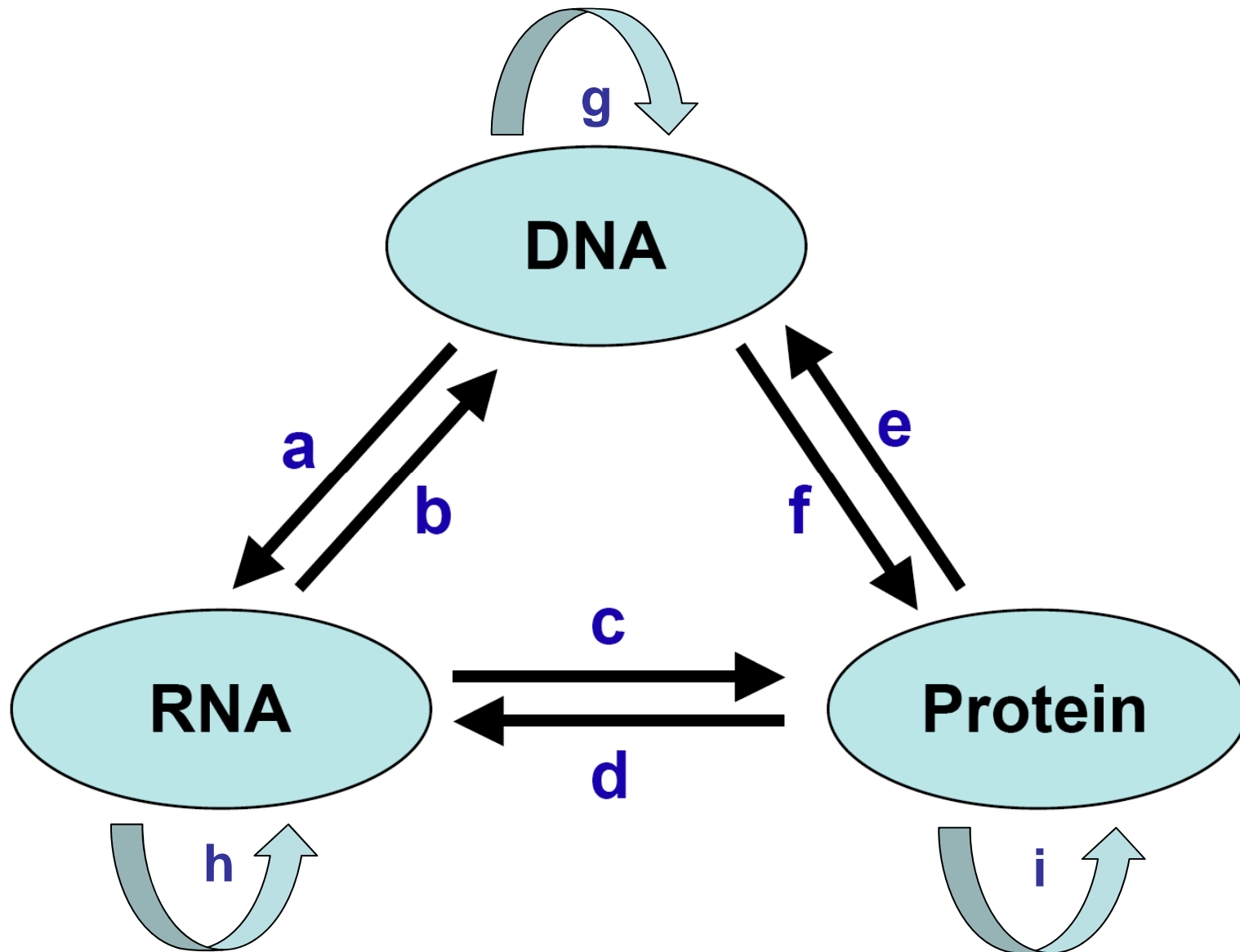


# DNA as Biological Information

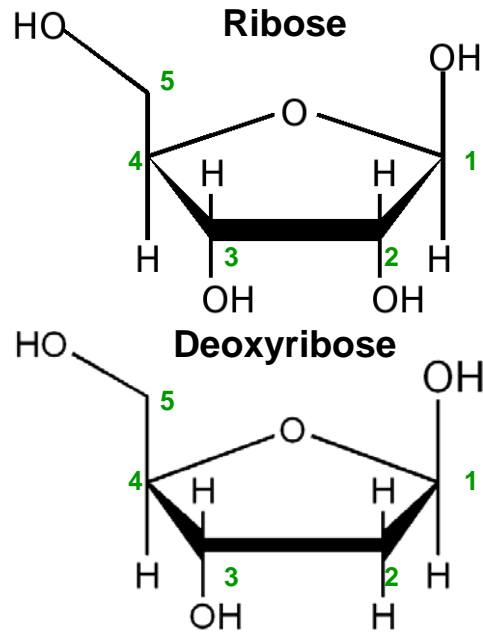
Rasmus Wernersson  
Henrik Nielsen

- Learning objectives
  - About Biological Information
  - A note about DNA sequencing techniques and DNA data
  - File formats used for biological data
  - Introduction to the GenBank database

# Information flow in biological systems



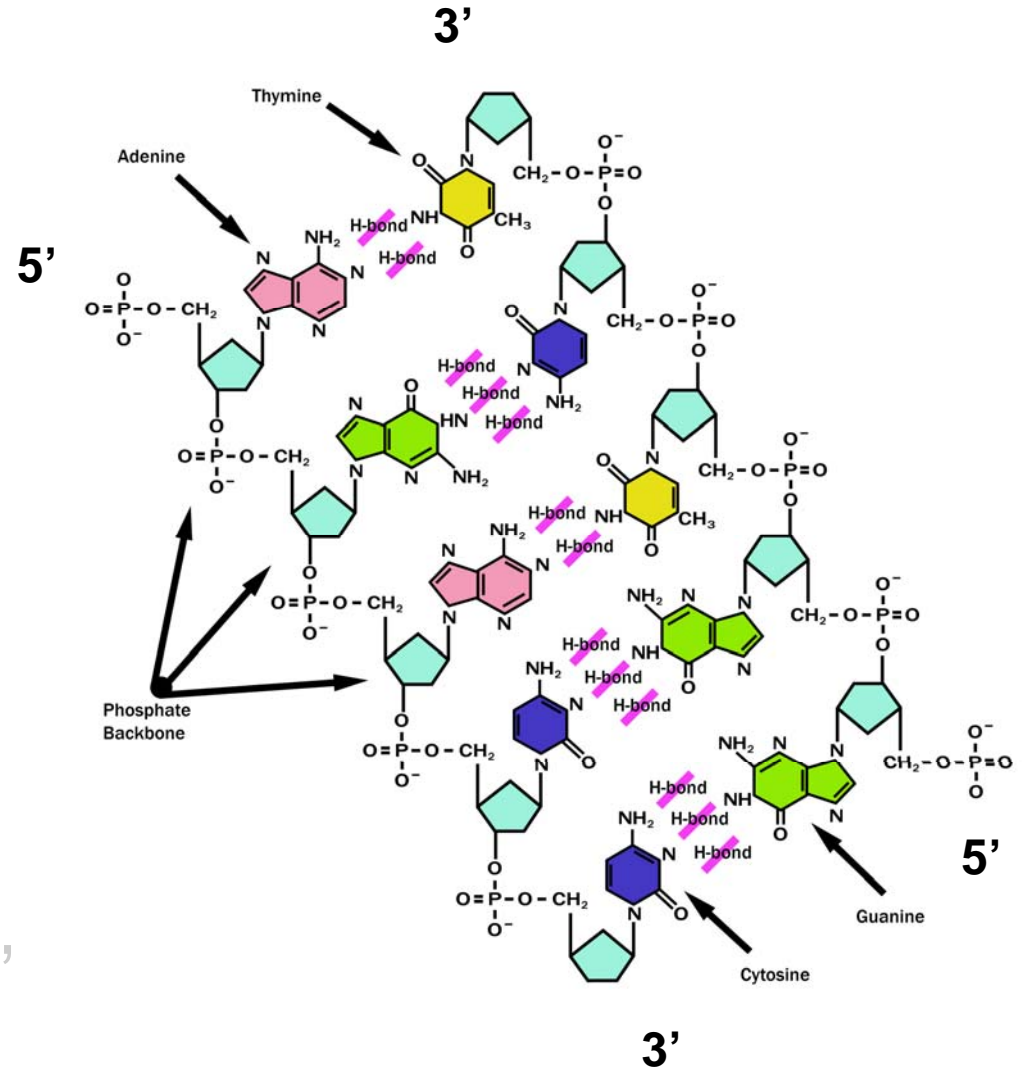
# DNA sequences = summary of information



5' AGCC 3'

3' TCGG 5'

5' ATGGCCAGGTAA 3'



Cycle 1

5'-CTAGATATGAACCTATAGGTACGGTGGCCATTCTATGTCTGATCCCGGTACTACCTACAGAA-3'  
|||||  
3'-GGGCCATGATGG-5'

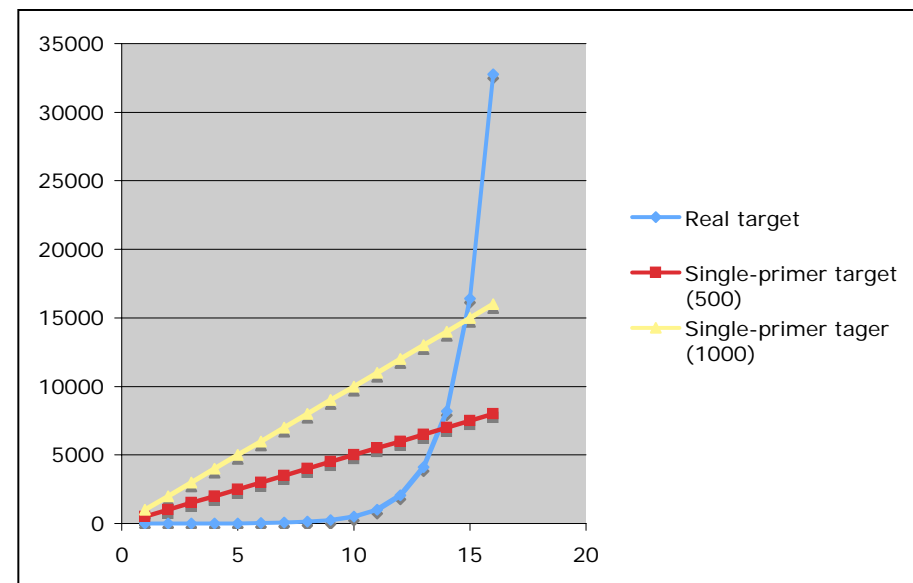
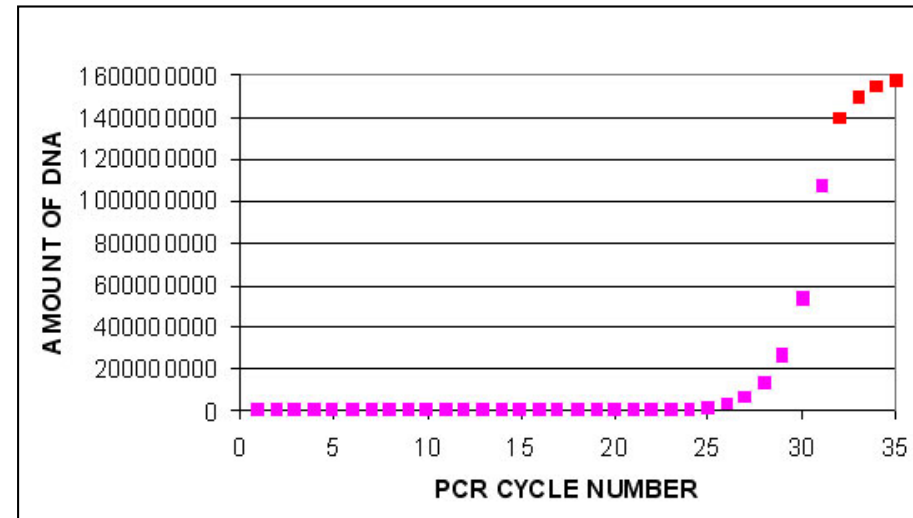
5'-ATGAACCTATAG-3'  
|||||  
3'-GATCTTATACTTTGGATATCCATGCCACCGGTAGATACAGACTAGGGCCATGATGGATGTCTT-5'

35  
cycles

**Melting**  
96°, 30 sec

**Annealing**  
~55°, 30 sec

**Extension**  
72°, 30 sec

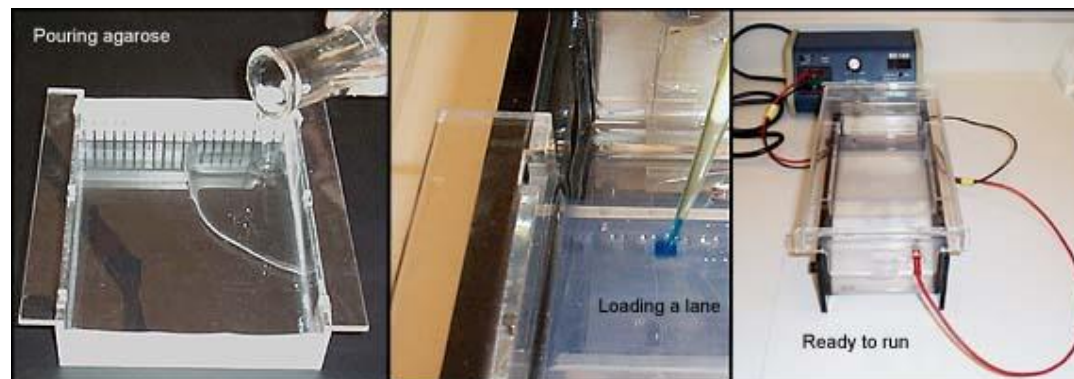
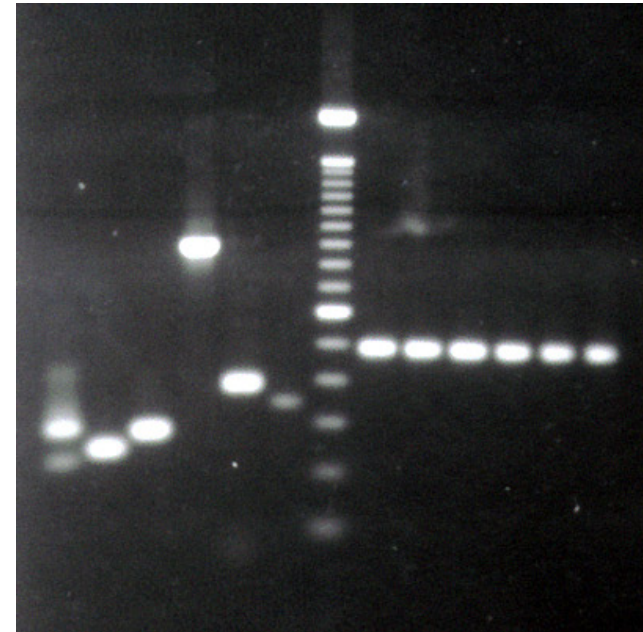


**Animation:** <http://www.people.virginia.edu/~rjh9u/pcranim.html>

**PCR graph:** <http://pathmicro.med.sc.edu/pcr/realtime-home.htm>

# Gel electrophoresis

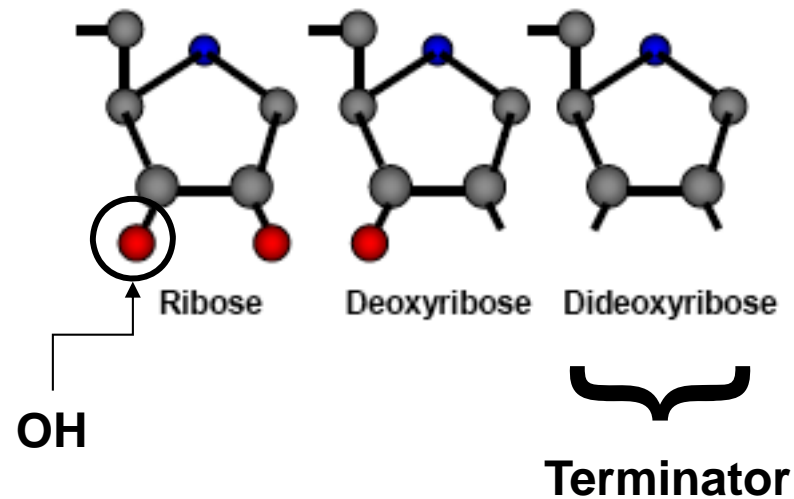
- DNA fragments are separated using gel electrophoresis
  - Typically 1% agarose
  - Colored with EtBr or ZybrGreen (glows in UV light).
  - A DNA "ladder" is used for identification of known DNA lengths.



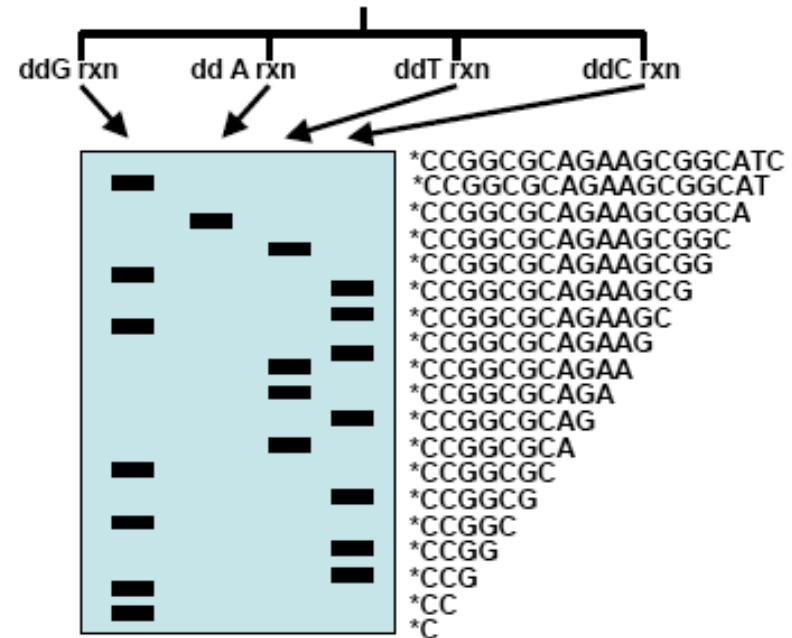
Gel picture: <http://www.pharmaceutical-technology.com/projects/roche/images/roche3.jpg>

PCR setup: <http://arbl.cvmbs.colostate.edu/hbooks/genetics/biotech/gels/agardna.html>

# The Sanger method of DNA sequencing



5' pCpCpGpGpCpGpCpApGpApApGpCpGpGpCpApTpCpApGpCpApApA 3'

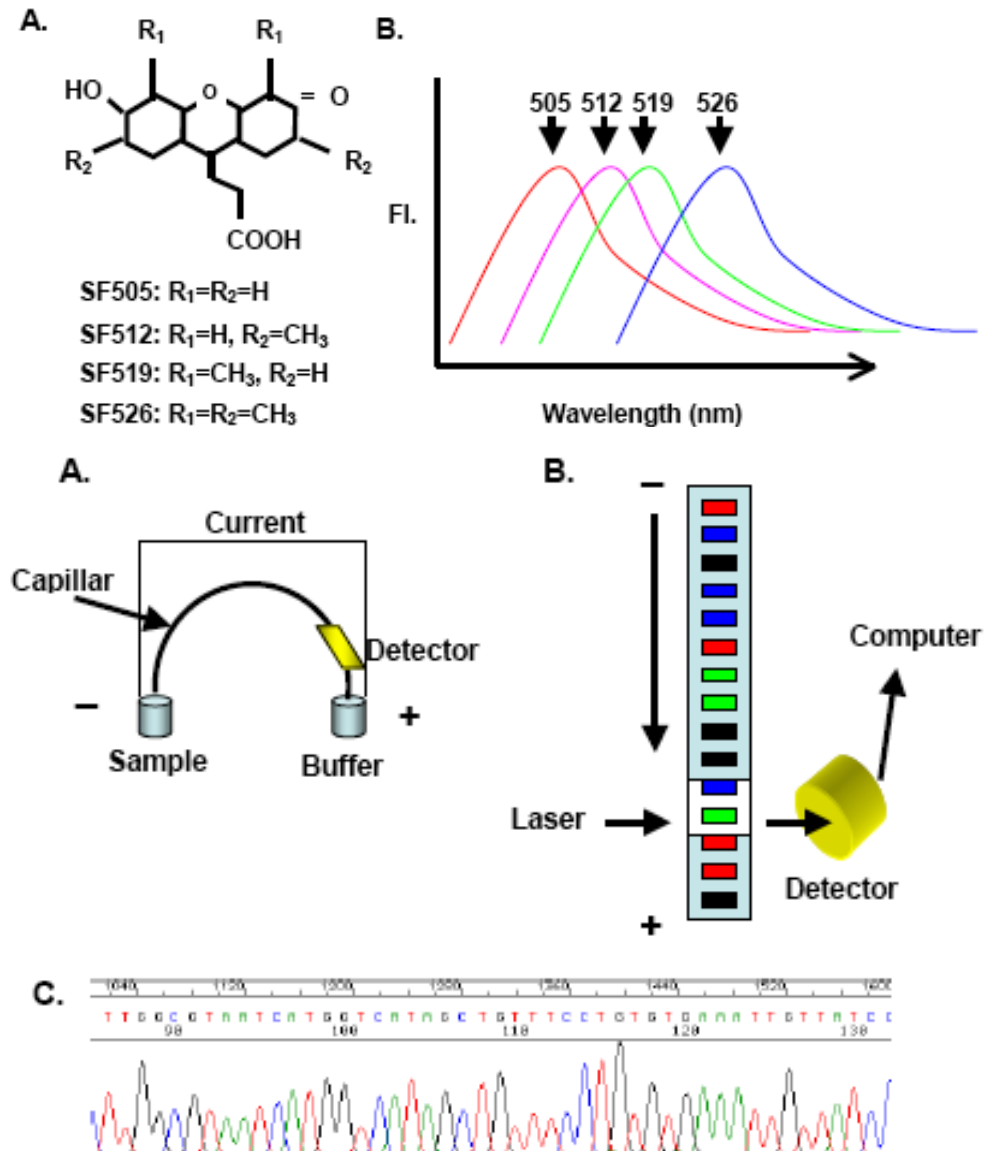


X-ray sequencing gel



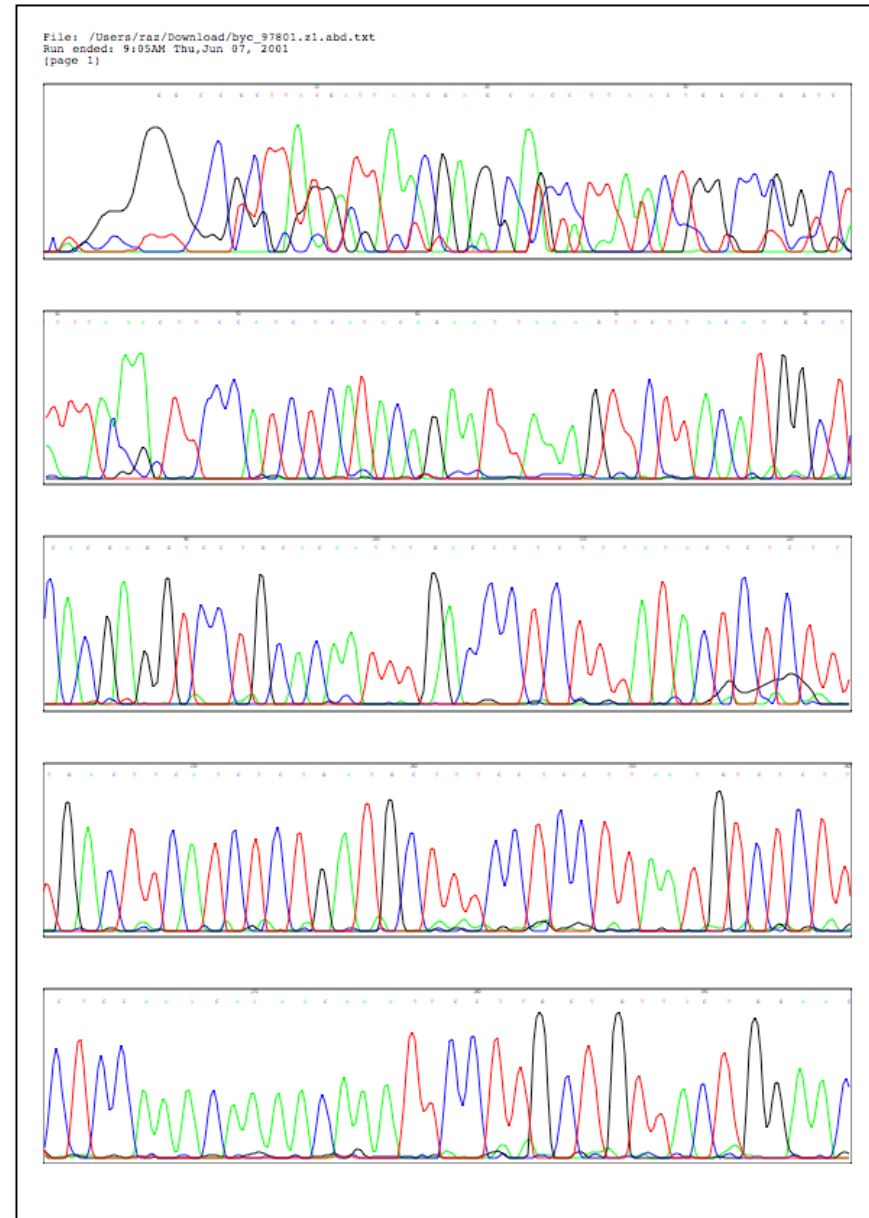
# Automated sequencing

- The major break-through of sequencing has happened through *automation*.
- Fluorescent dyes.
- Laser based scanning.
- Capillary electrophoresis
- Computer based base-calling and assembly.

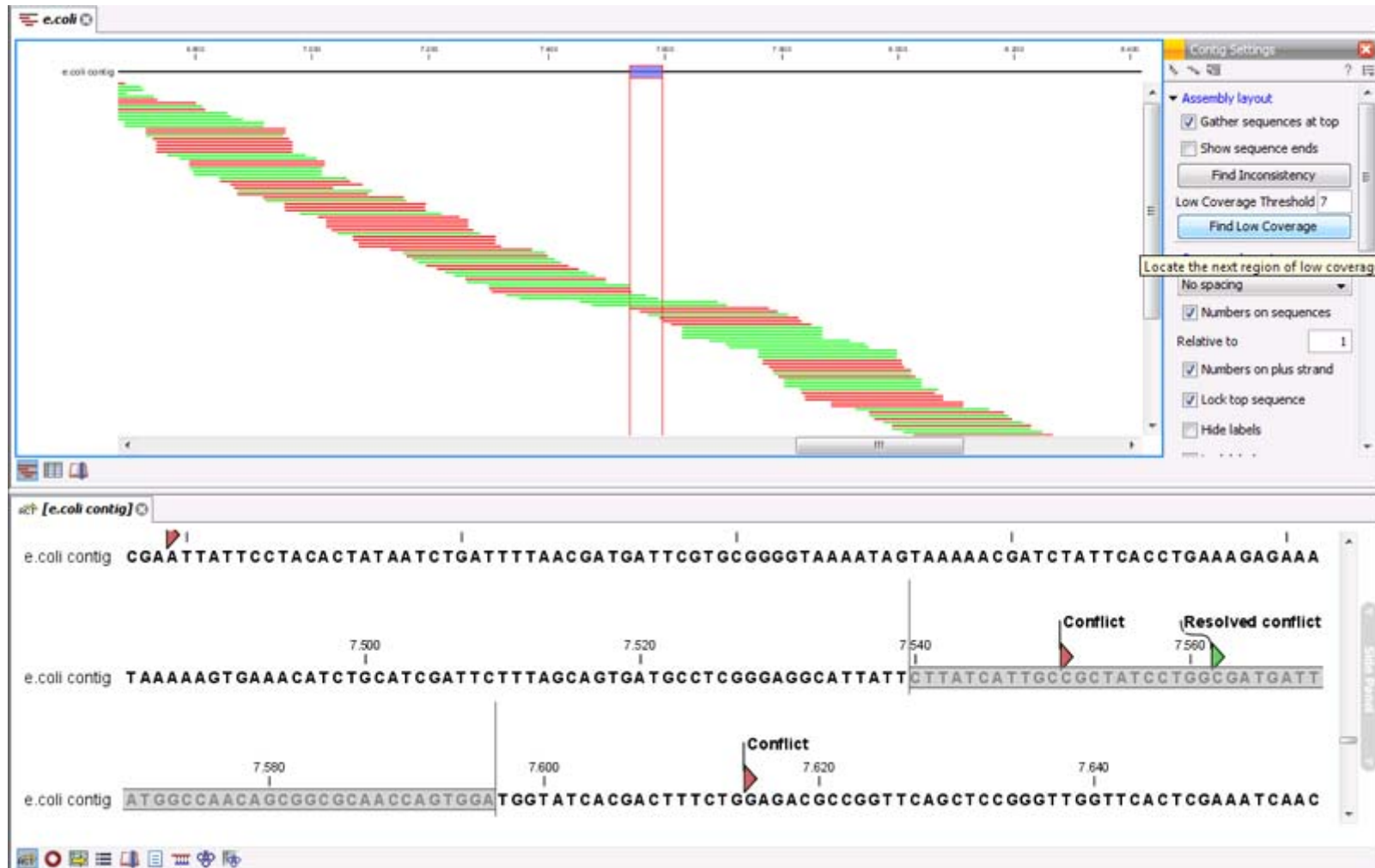


# Handout exercise: "base-calling"

- Handout:  
Chromatogram
- Groups of 2-3.
- Tasks:
  - Identify "difficult" regions
  - Identify likely errors
  - Try to estimate the best interval to use



# Sequence read mapping

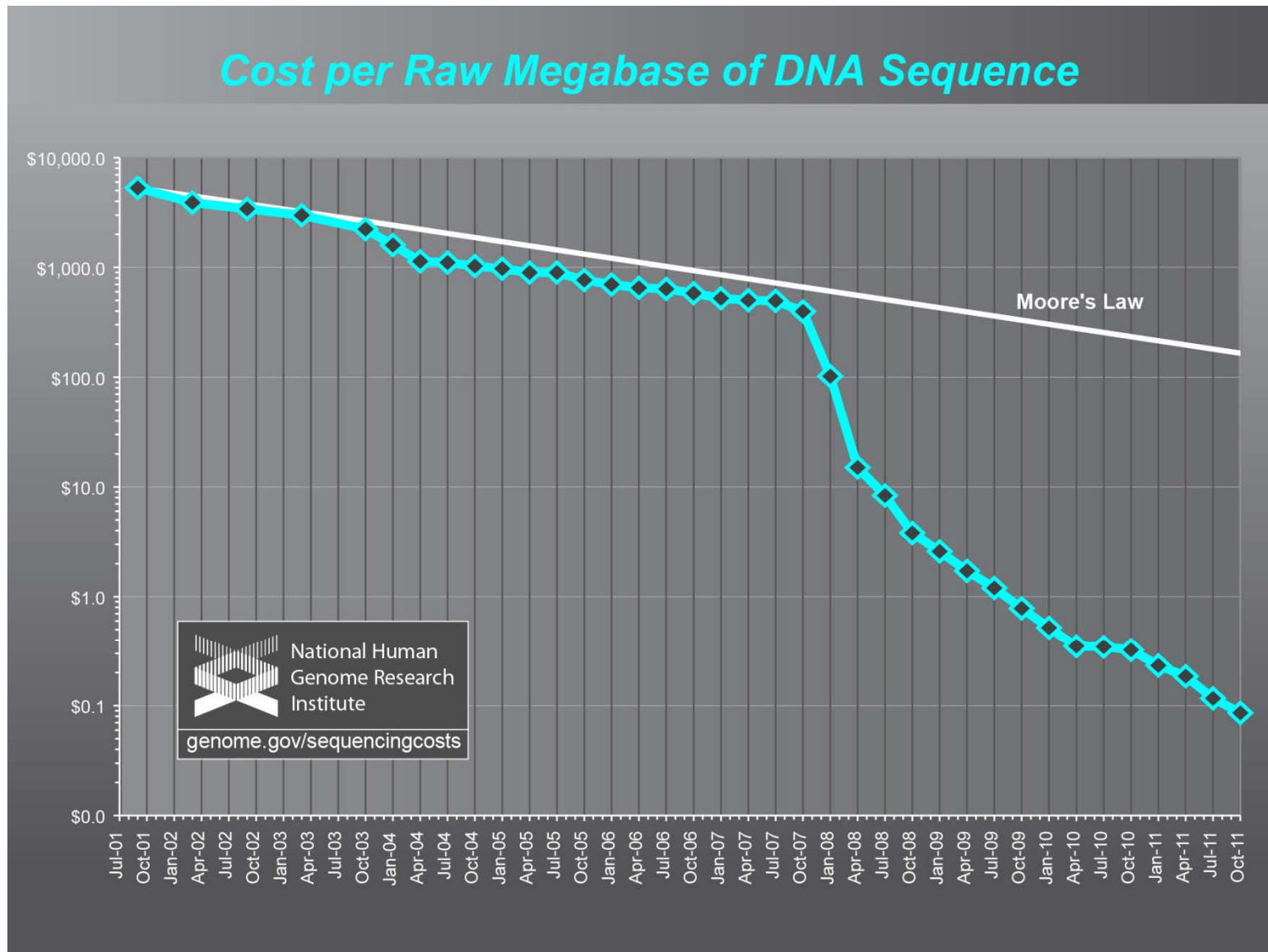


# DNA sequencing — history

---

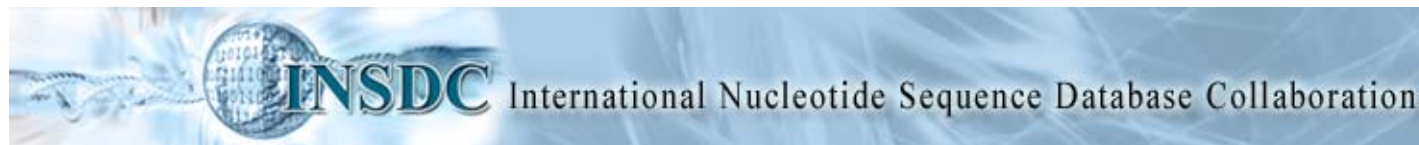
- 1972** Recombinant DNA technology [Paul Berg].
- 1976** The first sequenced genome, the bacteriophage MS2 (actually, RNA) [Walter Fiers *et al.*]
- 1977** DNA sequencing by chemical cleavage [Allan Maxam & Walter Gilbert]; DNA sequencing by enzymatic synthesis [Fred Sanger].
- 1982** *GenBank* is established.
- 1987** The first automatic sequencer, *Prism 373* [Applied Biosystems].
- 1990** *Human Genome Project* is launched.
- 1995** The first genome of a free-living organism, the bacterium *Haemophilus influenzae* (1.8 Mb) [The Institute for Genomic Research (TIGR)].
- 1996** The first genome of a eukaryote, Baker's yeast, *Saccharomyces cerevisiae* (12.1 Mb) [International consortium].
- 1998** The first genome of an animal, the nematode *Caenorhabditis elegans* (97Mb) [Sanger Center *et al.*].
- 2001** The first “drafts” of the Human genome (3Gb) [Human Genome Project Consortium (Nature, 15 Feb) + Celera (Science, 16 Feb)].
- Apr 15, 2012** *GenBank release 189* contains 151,824,421 sequences with a total of 139,266,481,398 nucleotides (the files take up 586 GB).

# Cost of sequencing



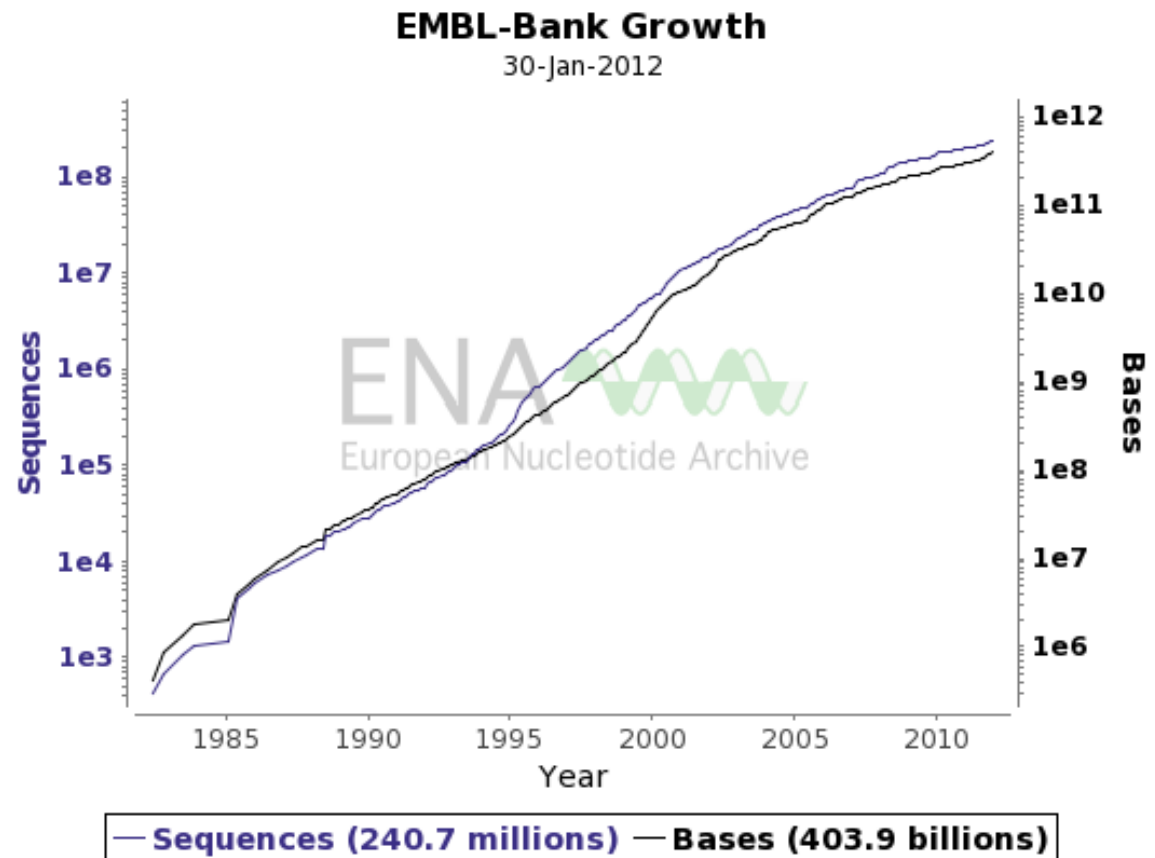
# Nucleotide databases

- **GenBank**, <http://www.ncbi.nlm.nih.gov/Genbank/>
  - National Center for Biotechnology Information (NCBI), National Library of Medicine (NLM), National Institutes of Health (NIH), USA
  - Established in 1982.
  - **EMBL**, <http://www.ebi.ac.uk/embl/>
  - European Bioinformatics Institute (EBI), England
  - Established in 1980 by the European Molecular Biology Laboratory, Heidelberg, Germany
  - Now part of **ENA**, the European Nucleotide Archive, <http://www.ebi.ac.uk/ena/>
  - **DDBJ**, <http://www.ddbj.nig.ac.jp/>
  - National Institute of Genetics, Japan
- 
- *Together they form*
  - International Nucleotide Sequence Database Collaboration, <http://www.insdc.org/>



# Nucleotide database growth

- Growth is roughly exponential
- But doubling time has increased from ~20 months (1990s) to ~50 months (2010)
- NB: The databases are public — no restrictions on the use of the data within.



## >alpha-D

```
ATGCTGACCGACTCTGACAAGAAGCTGGTCCTGCAGGTGTGGGAGAAGGTGATCCGCCAC
CCAGACTGTGGAGCCGAGGCCCTGGAGAGGTGCGGGCTGAGCTTGGGGAAACCATGGGCA
AGGGGGGCGACTGGGTGGGAGCCCTACAGGGCTGCTGGGGGTTGTTTCGGCTGGGGGTCAG
CACTGACCATCCCGCTCCCGCAGCTGTTACACCTACCCCCAGACCAAGACCTACTTCC
CCCACTTCGACTTGCACCATGGCTCCGACCAGGTCCGCAACCACGGCAAGAAGGTGTTGG
CCGCCTTGGGCAACGCTGTCAAGAGCCTGGGCAACCTCAGCCAAGCCCTGTCTGACCTCA
GCGACCTGCATGCCTACAACCTGCGTGTGACCCCTGTCAACTTCAAGGCAGGCGGGGGAC
GGGGGTCAGGGGCCGGGGAGTTGGGGGCCAGGGACCTGGTTGGGGATCCGGGGCCATGCC
GGCGGTACTGAGCCCTGTTTTGCCTTGCAGCTGCTGGCGCAGTGCTTCCACGTGGTGCTG
GCCACACACCTGGGCAACGACTACACCCCGGAGGCACATGCTGCCTTCGACAAGTTCCTG
TCGGCTGTGTGCACCGTGCTGGCCGAGAAGTACAGATAA
```

## >alpha-A

```
ATGGTGCTGTCTGCCAACGACAAGAGCAACGTGAAGGCCGTCTTCGGCAAAATCGGCGGC
CAGGCCGGTGACTTGGGTGGTGAAGCCCTGGAGAGGTATGTGGTCATCCGTCATTACCCC
ATCTCTTGTCTGTCTGTGACTCCATCCCATCTGCCCCCATACTCTCCCCATCCATAACTG
TCCCTGTTCTATGTGGCCCTGGCTCTGTCTCATCTGTCCCCAACTGTCCCTGATTGCCTC
TGTCCCCCAGGTTGTTTCATCACCTACCCCCAGACCAAGACCTACTTCCCCCACTTCGACC
TGTCACATGGCTCCGCTCAGATCAAGGGGCACGGCAAGAAGGTGGCGGAGGCACTGGTTG
AGGCTGCCAACCACATCGATGACATCGCTGGTGCCCTCTCCAAGCTGAGCGACCTCCACG
CCCAAAGCTCCGTGTGGACCCCGTCAACTTCAAAGTGAGCATCTGGGAAGGGGTGACCA
GTCTGGCTCCCCCTCCTGCACACACCTCTGGCTACCCCTCACCTACCCCCCTTGCTCACC
ATCTCCTTTTGCCTTTTCACTGCTGGGTCACTGCTTCCTGGTGGTCGTGGCCGTCCACTT
CCCCTCTCTCCTGACCCCGGAGGTCCATGCTTCCCTGGACAAGTTCGTGTGTGCCGTGGG
CACCGTCCTTACTGCCAAGTACCGTTAA
```



## GenBank format

```
LOCUS       CMO20202                1189 bp    DNA     linear   VRC 18-APR-2009
DEFINITION  Cairina moschata (duck) gene for alpha-D globin.
ACCESSION   X01831
VERSION     X01831.1  Gi162724
KEYWORDS    alpha-globin; globin;
SOURCE      Cairina moschata (Muscovy duck)
  ORGANISM  Cairina moschata
            Eukaryota; Neornithae; Chordata; Craniata; Vertebrata; Euteleostomi;
            Archosauria; Aves; Neognathae; Anseriformes; Anatidae; Cairina.
REFERENCE   1  (bases 1 to 1189)
AUTHORS     Erdil, C. and Wieseling, J.
TITLE       The primary structure of the duck alpha D-globin gene: an unusual
            5' splice junction sequence
JOURNAL     EMBO J. 2 (8), 1339-1343 (1983)
PUBMED      16872328
COMMENT     Data kindly reviewed (13-NOV-1983) by J. Wieseling.
FEATURES             Source: Source
             source
             ..1189
             /organism="Cairina moschata"
             /mol_type="genomic DNA"
             /db_xref="taxon:9855"

             CAAT_signal
             ..24
             TATA_signal
             ..73
             precursor_mRNA
             101..1114
             /note="primary transcripts"
             exon
             101..234
             /number=1
             CDS
             join(143..234,387..591,939..1067)
             /codon_start=1
             /product="alpha D-globin"
             /protein_id="CA025965.2"
             /db_xref="GI:4455876"
             /db_xref="Gene:PO2003"
             /db_xref="InterPro:IPR000971"
             /db_xref="InterPro:IPR002338"
             /db_xref="InterPro:IPR002340"
             /db_xref="InterPro:IPR009090"
             /db_xref="UniProtKB:Prot:PO2003"
             /translation="MTAEKPKLVQVNRVAVRQERFQIRALQNPAYVTKTTF
             NPLRPRGQVNRVAVRQERFQIRALQNPAYVTKTTFNPLRPRGQVNRVAVRQERFQIRALQNPAYVTKTTF
             NPLRPRGQVNRVAVRQERFQIRALQNPAYVTKTTFNPLRPRGQVNRVAVRQERFQIRALQNPAYVTKTTF"

             repeat_region
             227..246
             /note="direct repeat 1"
             intron
             235..386
             /number=1
             repeat_region
             387..399
             /note="direct repeat 1"
             exon
             387..391
             /number=2
             intron
             512..539
             /number=2
             exon
             940..1114
             /number=3
             polyA_signal
             1095..1102
             polyA_signal
             1114

ORIGIN
1  ctgagtgagg  taagcccttc  caaccctcca  cgtgataaag  ataagggcag  ggggggagc
61  cagggtgata  taagagctgg  gggggggggg  tggcttcaac  aagaagaaac  gtagctgac
121  agcgtgcaac  ggcgtgagcg  cagatgtgat  cgcagagpac  aagaagctca  tctgtcaggt
181  ttggagagag  gtcagtgagg  aacaggggag  atagagagat  gaagctgtgt  agagataga
241  gttgagggca  ggggggacct  acaggttgag  cagcagggag  caggagccat  gacggcggg
301  ttgggtggga  cccagagagc  caagggggag  gggctgagat  gggtaagaca  gacgggacac
361  aaaaatgagt  gggctgagat  gggagagaga  tctctggagt  aacacagagc  caagagatca
421  ttctctctct  tggactgaga  tctggggggt  gaaacagctc  gtagcagtag  caagaagagt
481  ggggtggtcc  ttgggagatg  cgtcagagag  ctggagcaac  taagcagagg  cctgtgtgag
541  ctacagaaac  tgaatgctta  caactgtgat  gtagcttgat  taacttttaa  ggaagcggg
601  gactaggtgt  cttaggttga  ggggttggag  gttgaggtgt  gacggcttga  ggggtttagg
661  ggtttgagtt  ttttgaggta  ttggagttat  ggggtttagg  gggcaggtta  atgtgttttt
721  ggttcaagg  gttttgggg  ttgagagaga  gacagagagg  gttgggattg  taatttgagt
781  ttggggcaga  ggttgagatt  gttttgaga  ttgggtttgt  gacgggttga  ggggttgggt
841  ggggttgaac  ggggtttagg  ggggtttagg  ggggtttagg  ggggtttagg  ggggtttagg
901  ttggagagag  ggtttagtag  cctgtgtttg  cctgtgtgat  gctgtgtgat  tgtgtgtgat
961  ttgtgtgtgt  ggcacacagt  ggcacacagt  aacagcctga  gatgtgtgat  gttttgtgat
1021  agttctgtgt  ggtgtgtgat  ggtgtgtgat  gtcacacagt  aagatgtgat  aagatgtgat
1081  cctgtgtgat  ttcacacagt  aacacacagt  cagagctgtg  tgtgtgtgat  tgtgtgtgat
1141  ggggtttagg  ggttttagg  ggttttagg  ggttttagg  ggttttagg  ggttttagg
```

Header

Indeholder information ang. Organisme, publikation, Accession ID mm.

FEATURE blok

Indeholder en beskrivelse af forskellige elementer i DNA sekvensen.

CDS: Coding Sequence. Indeholder koordinater på den protein kodende del af et gen. Bemærk de tre intervaller.

ORIGIN blok

Indeholder selve DNA sekvensen.

- Originates from the GenBank database.
- Contains both a DNA sequence and annotation of feature (e.g. Location of genes).

# GenBank format - HEADER

---

LOCUS CMGLOAD 1185 bp DNA linear VRT 18-APR-2005  
 DEFINITION Cairina moschata (duck) gene for alpha-D globin.  
 ACCESSION X01831  
 VERSION X01831.1 GI:62724  
 KEYWORDS alpha-globin; globin.  
 SOURCE Cairina moschata (Muscovy duck)  
 ORGANISM Cairina moschata  
 Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;  
 Archosauria; Aves; Neognathae; Anseriformes; Anatidae; Cairina.  
 REFERENCE 1 (bases 1 to 1185)  
 AUTHORS Erbil,C. and Niessing,J.  
 TITLE The primary structure of the duck alpha D-globin gene: an unusual  
 5' splice junction sequence  
 JOURNAL EMBO J. 2 (8), 1339-1343 (1983)  
 PUBMED 10872328  
 COMMENT Data kindly reviewed (13-NOV-1985) by J. Niessing.

# GenBank format - ORIGIN section

## ORIGIN

```

1  ctgcgtggcc tcagccctc caccctcca cgctgataag ataaggccag ggcgggagcg
61  caggggtgcta taagagctcg gcccgcggg tgtctccacc acagaaacc gtcagttgcc
121 agcctgccac gccgctgccg ccatgctgac cgccgaggac aagaagctca tcgtgcaggt
181 gtgggagaag gtggctggcc accaggagga attcggaagt gaagctctgc agaggtgtgg
241 gctgggcca gggggcactc acaggggtgg cagcaggag caggagccct gcagcgggtg
301 tgggctggga ccagagcgc cacggggtgc gggctgagat gggcaaagca gcagggcacc
361 aaaactgact ggcctcgctc cggcaggatg ttctcgctt accccagac caagacctac
421 tccccccact tcgacctgca tcccggtctt gaacagggtc gtggccatgg caagaaagtg
481 gcggctgccc tgggcaatgc cgtgaagagc ctggacaacc tcagccaggc cctgtctgag
541 ctcagcaacc tgcattgcta caacctgcgt gttgacctg tcaacttcaa ggcaagcggg
601 gactagggtc cttgggtctg ggggtctgag ggtgtggggt gcagggtctg ggggtccagg
661 ggtctgagtt tcctggggtc tggcagtcct gggggctgag ggccagggtc ctgtggtctt
721 gggtagcagg gtcctggggg ccagcagcca gacagcaggg gctgggattg catctgggat
781 gtgggcccaga ggctgggatt gtgtttggaa tgggagctgg gcaggggcta gggccagggt
841 gggggactca gggcctcagg gggactcggg gggggactga gggagactca gggccatctg
901 tccggagcag gggactaag ccctggtttg ccttgagct gctggcacag tgcttccagg
961 tgggtgctggc cgcacacctg ggcaaagact acagccccga gatgcatgct gcctttgaca
1021 agttcttgct cgccgtggct gccgtgctgg ctgaaaagta cagatgagcc actgcctgca
1081 cccttgacac ttcaataaag acaccattac cacagctctg tgtctgtgtg tgctgggact
1141 gggcatcggg ggtcccaggg agggctgggt tgcttccaca catcc

```

//

# GenBank format - FEATURE section

```

FEATURES                     Location/Qualifiers
    source                    1..1185
                              /organism="Cairina moschata"
                              /mol_type="genomic DNA"
                              /db_xref="taxon:8855"
    CAAT_signal               20..24
    TATA_signal               69..73
    precursor_RNA             101..1114
                              /note="primary transcript"
    exon                      101..234
                              /number=1
    CDS                       join(143..234,387..591,939..1067)
                              /codon_start=1
                              /product="alpha D-globin"
                              /protein_id="CAA25966.2"
                              /db_xref="GI:4455876"
                              /db_xref="GOA:P02003"
                              /db_xref="InterPro:IPR000971"
                              /db_xref="InterPro:IPR002338"
                              /db_xref="InterPro:IPR002340"
                              /db_xref="InterPro:IPR009050"
                              /db_xref="UniProt/Swiss-Prot:P02003"
                              /translation="MLTAEDKKLIVQVWEKVAGHQEEFGSEALQRMFLAYPQTKTYFP
HFDLHPGSEQVRGHGKKVAAALGNAVKSLDNLSQLSELNLHAYNLRVDPVNFKLLA
QCFQVVLAHLGKDYSPEMHAADFDFLSAVAAVLAEKYR"
    repeat_region             227..246
                              /note="direct repeat 1"
    intron                    235..386
                              /number=1
    repeat_region             289..309
                              /note="direct repeat 1"
    exon                      387..591
                              /number=2
    intron                    592..939
                              /number=2
    exon                      940..1114
                              /number=3
    polyA_signal              1095..1100
    polyA_signal              1114

```

# Exercise: GenBank

- Work in groups of 2-3 people.
- The exercise guide is linked from the course programme.
- Read the guide carefully - it contains a lot of information about GenBank.

The screenshot shows the Entrez Nucleotide database homepage. The browser address bar displays the URL: <http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?CMD=>. The page features a navigation bar with links to All Databases, PubMed, Nucleotide, Protein, Genome, Structure, PMC, Taxonomy, and Books. A search bar is prominently displayed with the text "Search Nucleotide for" and buttons for "Go" and "Clear". Below the search bar, there are tabs for "Limits", "Preview/Index", "History", "Clipboard", and "Details". The main content area includes a welcome message about the Entrez Nucleotides database, a section for the Human Genome with a link to explore human genome resources, and a section for building the human genome with a link to the Human Genome Reference DNA Sequence. At the bottom, there is a section for the Homo sapiens genome view, showing a map of the human genome with chromosomes 1 through 22, X, and Y.